CHARACTERIZATION OF EXTREME/RARE EVENTS AND DATA ANALYSIS

Zhigang Wei[1] **Kamran Nikbin**[2] **Limin Luo**[1] **Litang Gao**[1]

[1]*Tenneco Inc., Grass Lake, Michigan, USA*
[2]*Department of Mechanical Engineering, Imperial College London, UK*

ABSTRACT

In this paper the fundamental probability concepts and two commonly used probabilistic distribution functions, i.e. the Weibull for spectrum events and the Pareto for extreme/rare events, are described first. An events quadrant is subsequently established based on the commonality/rarity and impact/effect of the probabilistic events. Level of measurement, which is the key for a quantitative measurement, is also discussed based on the framework of probability diagram. Four case studies, i.e. vehicle road test score, the fatigue life distribution of a metallic material, the city population distribution in 3 countries, and the earthquake distribution worldwide and in the USA, are provided to demonstrate the probabilistic approaches on events characterization and data analysis. Two possible holistic mechanisms, i.e. equilibrium and evolution based mechanisms, for empirical distributions are provided in the Appendices.

1. INTRODUCTION

A probabilistic distribution function roughly consists of two parts: the middle and the tails. Based on the contribution of the tails to the overall damage or impact, probabilistic events can be divided into two categories: (1) extreme/rare events, and (2) spectrum events [1]. The damage caused by the extreme/rare events, such as 100-year flood or mega-earthquake or nuclear accident etc. is controlled by the tails. By contrast, the damage caused by the constituents in the spectrum events is comparable, oftentimes, dominated by the mean behaviors. The two categories of the probabilistic events are different in nature. How to characterize and categorize these probabilistic events and subsequently analyze the data are critical to probabilistic risk and safety management [2].

The traditional continuous two-parameter normal, lognormal, and Weibull distributions are often used to describe the probabilistic distribution of spectrum events such as fatigue life. Since these functions have either infinite lower bound or infinite upper bound, or both, they are not capable to accurately describe the tail behaviors though they are often used in practice. To properly overcome the shortcomings of the two-parameter distribution functions, multiple-parameter distribution function which considers the tail behaviors have been developed [1]. For the extreme events which are controlled by the tail behavior, the conventional distribution functions do not work well. Other distribution functions such as the Pareto power law and exponential functions are more appropriate. The difference and similarity between the functions for spectrum events, such as Weibull, and the functions for extreme/rare events, such as Pareto, and the relationship between them is of significant importance in fundamentally understanding of the probabilistic events and in practical applications.

It is well known that many phenomena in both the natural and social sciences have power law statistics (Pareto distribution). The phenomena include city sizes, incomes, word frequencies, earthquake magnitudes, and many other natural and manmade engineering events [3]. A power-law distribution implies that small occurrences are extremely common, whereas large instances are extremely rare. However, it has been found that despite the fact that a power law models the tails of the empirical distribution well, the largest events are significantly outlying, meaning much larger or smaller than what would be expected under the power law. Such events are interpreted as "Dragon Kings" as they indicate a departure from the generic process underlying the power law [4, 5]. The extreme/rare events are difficult to understand in their nature and are traditionally treated as "Black Swan" phenomena [6] meaning they rarely happen, with significant consequences when they happen, but are unpredictable. However, since these phenomena bear significant consequences, deep understandings of these "Dragon Kings" behaviors and substantial development of the related predictive tools would be highly demanded [4, 5].

In this paper, the general probabilistic description of events is given first, with emphasis on the Weibull and Pareto power law distributions, their relationships, similarities and differences. Level of measurement, which is fundamental in quantitatively assessment of probabilistic events is also reviewed and discussed. Events categorization based on commonality/rarity and impact/effect are described subsequently and an events quadrant is proposed for ease of events categorization. Finally, several examples are provided to demonstrate the concepts and ideas developed in this paper.

2. CHARACTERIZATION OF PROBABILISTIC EVENTS

2.1 Description of a probabilistic event and level of measurement

Probabilistic approaches are more appropriate than the deterministic approaches to describe the natural and engineering world, which contains a wide variety of uncertainties. The traditional deterministic approaches produce exactly the same result no matter how many times the event is repeated under the same condition. Probability is a measure of likelihood that an event will occur, and it is quantified as a number between 0 and 1. The higher the probability of an event, the more certain that the event will occur. 0 indicates impossibility and 1 indicates certainty. The probabilistic behavior of an event can be described by several parameters as schematically shown in the probability-variable ($P-x$) diagram, Fig.1. In Fig.1 x along the horizontal axis represents the variable of interest while P along the vertical axis represents the cumulated distribution function (CDF), which relates the probabilistic density function (PDF) $p(x)$ through $P(x)=\int_a^x p(x)$ and $P(b)=\int_a^b p(x)=1$. The lower limit a and the upper limit b are finite values for most of the real problems. The infinites $a=-\infty$ and $b=\infty$ are often a mathematical idealization.

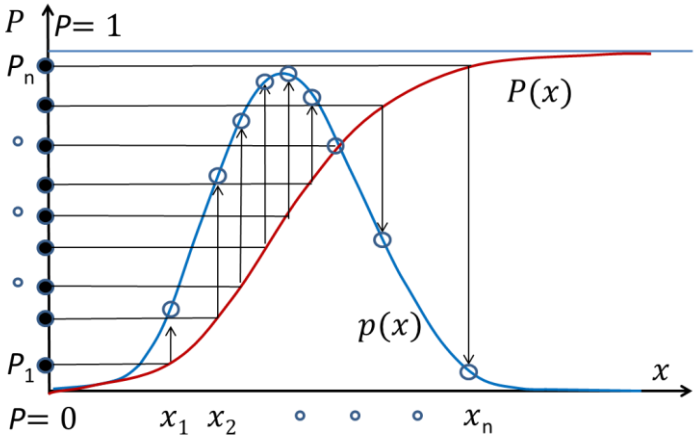


Fig.1 Schematic of relationship among several probabilistic measures.

For a sequence of n independent and random tests, there will be a corresponding discrete sequence of values of x, say x_1, x_2,… x_n (ranked in an ascending order). No matter what the $p(x)$ looks like, $P(x)$ can always be estimated based on the ranked data. The accuracy of the estimation depends on the sample size and the estimation model. For a given reasonably accurate model, the larger the sample size, the more accurate the estimated $P(x)$. The simplest estimation of $P(x)$ is i/n [7], which obviously has some issues. For example, when $i=0$ and $i=n$, the formula results in $P(x_1)=1/n$ and $P(x_n)=1$, which contradicts the fact that any randomly generated data can be close to but never reach $P=1$. Similarly, several other simple formulae such as $(i-1)/n$, $(i-1/2)/n$, and $i/(n+1)$ have similar issues [7]. Based on the cumulative binomial distribution and median rank (MR) formula, the so called Benard's approximation, Eq.(1), has been developed and now widely used in industry to estimate $P(x)$ [7].

$$P(x_i)\approx MR=\frac{i-0.3}{n+0.4} \tag{1}$$

where n is the total number of tests and i is the rank order number. It should be noted that the randomly generated data as projected onto the vertical axis P is uniformly distributed so that the binomial distribution can be assumed regardless of the information about $p(x)$. It should be emphasized here that the generation process of uniformly distributed random number is also the first step in Monte Carlo simulation for generating random number for any specific distributions, such as Weibull or Lognormal. The working mechanism of creating a probability plot using the median rank approach is schematically illustrated in the $P-x$ diagram shown Fig.1. The distribution of $p(x)$ is determined by the location of the data on x axis, which are obtained by vertically projecting the data located on the $P-x$ curve, which are initially projected horizontally from the uniformly generated data on the P axis. Corresponding to the uniform distributed data points on the vertical axis P, the

intervals between the adjacent limits on x axis are unequally spaced with the dense segments near the peak of the $p(x)$ curve. The segments become coarser as the distance from the peak goes further from the peak.

The exact forms of $P(x)$ and $p(x)$ also depend upon the level (unit) and the scale of the x axis. Level of measurement or scale of measure is a classification that describes the nature of information within the numbers assigned to variables. Some physical quantities, such as cycles to failure, energy, city population etc., can be directly used. The use of unit in log or exponential form and other objective and subjective measurements are also very common. Oftentimes, classification with 3-, 4-, 5-, 10-, 12-, 100- level measures are frequently used for ease of communication and simplicity. For example, Beaufort wind force scale (0-12) is an empirical measure that relates wind speed to observed conditions at sea or on land. The wind speed on a Beaufort scale (B) is based on the empirical relationship: $V = 0.836B^{3/2}$ m/s. Richter magnitude scale for earthquake assigns a magnitude number to quantify the energy or moment released by an earthquake. Richter scale is a base-10 logarithmic scale, which defines magnitude as the logarithm of the ratio of the amplitude of the seismic waves to an arbitrary, minor amplitude. Earthquakes are classified as moderate if their magnitude is in the range of 5-5.9, strong if the magnitude is in the range of 6-6.9, major if the magnitude is in the range of 7-7.9, and great if the magnitude is 8 or larger [8]. The decibel (dB) is also a logarithmic unit used to repress the ratio of two values of a physical quantity, often power or intensity. The Numeric Rating Scale (NRS-11) is an 11-point scale for patient self-reporting of pain. The FICO credit scores are designed to measure the risk of default by taking into account various factors in a person's financial history. The generic or classic FICO score is between 300 and 850. There are some other measures without numbering system but with color to represent severity level. For example, the US Homeland security advisory system is a color-code terrorism threat advisory scale: Green (Low), Blue (Guarded), Yellow (Elevated), Brown (High), and Red (Severe). A similar scale is also used in risk level characterization: Low (Green), Medium (Yellow), and High (Red). Clearly, different measures of the variables result in different distributions. Therefore, the empirical interpretation of the probabilistic behavior must be consistently based on a certain level of measurement. The commonly used Weibull distribution for spectrum events and the Pareto distribution for extreme/rare events are described below.

2.2 Weibull distribution functions

In many engineering applications, Weibull distribution functions are often used to describe a spectrum load, with which all of the loading spectra under normal operating conditions are evaluated so that the dominating contributions can be properly considered in product design and validation [1]. These loads include the loads experienced by the components in nuclear power plants, ground vehicles, and landing gears of aircrafts. A spectrum event, e.g. the fatigue life distribution from the left tail to the right tail, often shows a lower bound and an upper bound, meaning there is no failure below the lower bound while there is no survival beyond the upper bound. In these cases, the introduction of the threshold parameters, i.e. a floor parameter for the lower bound and the ceiling parameter for the upper bound, are necessary. The floor parameter is important in product validation and quality control while the ceiling parameter is directly related to lifecycle management, new product development, and profit generation for the manufactures.

The general five-parameter Weibull CDF [1] shown in Eq.(2) is such an example that can cover both the middle and the tails (both left and right) properly.

$$P(x) = 1 - \exp\left[-\frac{1}{\lambda} \frac{(x-a)^{\beta_1}}{(b-x)^{\beta_2}} \right]; 0 < a \le x \le b < \infty, \lambda, \beta_1, \beta_2 > 0 \tag{2}$$

In Eq.(2), a and b are the floor and ceiling parameters for characterizing the lower and the upper bound thresholds. β_1 and β_2 are two shape parameters respectively controlling the shape of the distribution at the left and the right tails. λ is the characteristic parameter controlling the location of the overall distribution function. The corresponding PDF can be easily obtained as $p(x) = dP(x)/dx$ but will not be elaborated here.

When $\beta_1 = \beta_2 = \beta$ and $\lambda = \eta^{\beta}$, Eq.(2) can be reduced to the four-parameter Weibull distribution (omitted here). When $b - x = 1$, the four-parameter Weibull distribution can be reduced to the three-parameter Weibull distribution, Eq.(3).

$$P(x) = 1 - \exp\left\{ -\left[\frac{(x-a)}{\eta} \right]^{\beta} \right\}; 0 < a \le x < \infty, \eta, \beta > 0 \tag{3}_1$$

$$p(x) = \left(\frac{\beta}{\eta} \right) \left(\frac{x-a}{\eta} \right)^{\beta-1} \exp\left[-\left(\frac{x-a}{\eta} \right)^{\beta} \right] \tag{3}_2$$

When $a = 0$, the three-parameter Weibull distribution can be further reduced to the conventional two-parameter (2P) Weibull distribution, Eq.(4).

$$P(x)=1-\exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right];0<x<\infty,\eta,\beta>0 \tag{4}_1$$

$$p(x)=\left(\frac{\beta}{\eta}\right)\left(\frac{x}{\eta}\right)^{\beta-1}\exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right] \tag{4}_2$$

It should be noted that similar kinds of four- and five-parameter Weibull distributions have been developed and applied to the strength distribution of glass [9], the fatigue growth rate [10], and tensile strength of optical fibers [11]. It is also noted that the two-parameter Weibull distribution can be reduced to the Rayleigh distribution when $\beta=2$, and to the exponential distribution when $\beta=1$.

2.3 Pareto (power law) distribution function

When dealing with crises and extremes, power law tails are the "normal" case [3]. The unique property of power law is that they are scale-invariant/self-similar/fractal. This property implies that all events, both large and small, are generated by the same mechanism. A continuous variable with a power-law distribution has a probability $f(x)dx$ of taking a value in the interval from x to $x+dx$, where $f(x)=Hx^{-\alpha}$ with $\alpha>0$ [3]. There must be some lowest value $x_{\min}$ at which the power law is obeyed, and for practical reasons, only the statistics of x above this value is considered. The constant in the power-law is given by the normalization requirement that $\int_{x_{\min}}^{\infty} f(x)dx=1$. Then, $H=(\alpha-1)x_{\min}^{\alpha-1}$ must be held. Finally, the Pareto PDF and CDF can be expressed in Eq.(5).

$$p(x)=\frac{\alpha-1}{x_{\min}}\left(\frac{x}{x_{\min}}\right)^{-\alpha} \tag{5}_1$$

$$P(x)=1-\left(\frac{x_{\min}}{x}\right)^{\alpha-1} \tag{5}_2$$

It should be noted that there is no cut-off as an upper bound in Eq.(5). Fortunately, in most cases the cutoff effects can be ignored for large sample sizes, in which the PDF value at $x_{\max}$ is extremely small.

The fit parameters of the Pareto distribution can be derived by applying the least squares method or the maximum likelihood method. However, results show that the maximum likelihood method provides a better data correlation than the least squares method [3]. The natural log likelihood function is shown in Eq.(6).

$$L(\alpha)=Ln\left[\prod_{i=1}^{n}\frac{\alpha-1}{x_{\min}}\left(\frac{x}{x_{\min}}\right)^{-\alpha}\right] \tag{6}$$

The maximum likelihood is found by differentiating $L(\alpha)$ in Eq.(6) with respect to parameter α, setting the result equal to zero. Upon rearrangement, this yields the estimator equation

$$\hat{\alpha}=1+n\left[\sum_{i=1}^{n}\ln\left(\frac{x_i}{x_{\min}}\right)\right]^{-1} \tag{7}$$

Where $x_i(i=1,...,n)$ are the n data points for $x_i\geq x_{\min}$. In practical situations $x_{\min}$ usually corresponds not to the smallest value of x measured but to the smallest for which the power-law behavior holds [3].

Identifying power-law behavior is important but the process can be tricky [3]. The standard strategy makes use of a histogram of a quantity with a power-law distribution appears as straight line when plotted on logarithmic scales. However, bin size selection is always an issue for this treatment and noise can be generated that hinders the data interpretation. Another, and in many ways a superior, method of plotting the data is to calculate a CDF [3]. Instead of plotting a histogram of the data, a plot of the CDF, called complementary CDF (C-CDF) here, has a value greater than or equal to x: $CP(x)=\int_{x}^{\infty} f(x')dx'$. If the distribution follows a power law, then

$$CP(x)=\frac{H}{\alpha-1}x^{-(\alpha-1)} \tag{8}$$

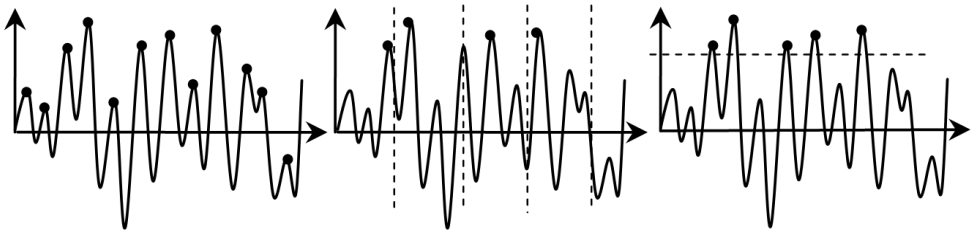
Thus, the C-CDF also follows a power law, but with an exponent of $\alpha-1$, indicating a straight line on logarithmic scales, but with a shallower slope if the power law distribution is held [3]. This can be used as a check if a distribution follows

the power law distribution. But notice that there is no need to bin the data at all to calculate $CP(x)$. By its definition, $CP(x)$ is well-defined for every value of x and so can be plotted as a perfectly normal function without binning. This avoids all questions about what sizes the bins should be [3].

Finally, it should be noted that the 2P-Weibull, Eq.(4), and the Pareto, Eq.(5), are related. For example if x is Pareto-distributed with minimum $x_{\min}$ and exponent α, then $y = \ln(x/x_{\min})$ is exponential distributed with rate parameter α. Equivalently, if y is exponentially distributed with rate α, then $x_{\min}\exp(y)$ is Pareto-distribution with minimum $x_{\min}$ and index α. Once again, an exponential distribution is a special case of 2P-Weibull when $\beta = 1$.

2.4 Characterization of extreme/rare events

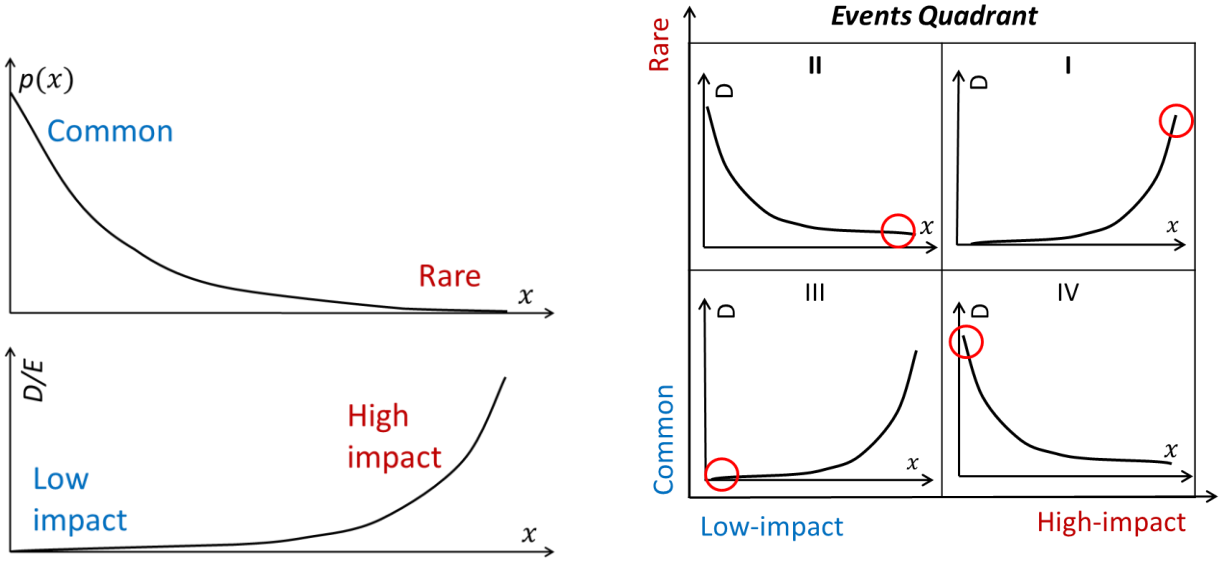
It is often desired to model extreme/rare events with a continuous distribution function. Generally, there are three ways of identifying extremes in a set of data, which could either be a time history or quasi-steady data: (1) peaks or valleys, Fig.2(a), (2) block maxima, where the maxima (or minima) in successive periods are selected, Fig.2 (b), and (3) peaks-over threshold, where the observations that exceed a given threshold are considered, Fig.2(c).



(a) (b) (c)

Fig.2 Schematic of extreme events (a) peaks, (b) block maxima, and (c) peak-over-threshold.

It has been found that the distribution of peaks follows the Rayleigh distribution when the random data follows Gaussian distribution. The block maxima follows the generalized extreme value distribution, and the peak-over-threshold follows the Pareto power law distribution, Eq.(5)[12]. It is also found that the extreme value theory provides a statistical justification for the emergence of power laws as limiting behavior for extreme fluctuations [13]. It should be noted that some events may be very common, such as low amplitude fatigue vibration as experienced by a vehicle, but their impact might not be remarkable, whereas some events might not be very common, but their consequence might be significant, such as a mega-earthquake. Based on the combination of the commonality/rarity and the impact/effect an event can be roughly categorized into one of the four events as shown in the events quadrant shown in Fig.3.



(a) (b)

Fig.3 Schematic of (a) a PDF function in exponential or power law form and the corresponding damage per event (D/E), and (b) the events quadrant based on the combination of commonality/rarity and impact/effect.

Fig.3(a) shows a monotone decreasing probabilistic distribution function, which could be either Pareto type or exponential type. Therefore, common events occur on the left side and the rare events occur on the right side. Corresponding to it the curve of the damage per event (D/E) curve is also shown in Fig.3(a). Assume that the damage is a monotone increasing function of the magnitude of the variable, which is often the case such as the load for fatigue and the earthquake magnitude. Clearly, the event that happens on the right side shows higher impact than that on the left side. Fig.3 (b) shows four typical combinations of the events in terms of total damage (D): Quadrant-I: Rare-High Impact, Quadrant-II: Rare-Low Impact, Quadrant-III: Common-Low Impact, and Quadrant-IV: Common-High Impact.

For fatigue failure, the event can be characterized by quantitatively evaluating a damage density function $g(S) = p(S)D(S)$ [1]. $p(S)$ is the probabilistic distribution function of stress range and $D(S)$ is the linearly accumulated damage per cycle $D = 1/N_f$. The expected total accumulated damage within time T is $E[D] = E[R]T\int_0^m g(S)dS = \frac{1}{C^m}E[R]T\int_0^\infty S^m p(S)dS$ for a fatigue S-N curve $S = CN_f^{-1/m}$. N_f is cycle to failure and $E[R]$ is rainflow counted cycle number in a unit time. For the Pareto power-law distribution shown in Eq.$(5)_1$, the damage density as a function of stress range can be expressed as $g(S) = \frac{\alpha - 1}{C^m S_{\min}^{1-\alpha}} S^{m-\alpha}$ [1]. Clearly, there are simply three scenarios for the Pareto power-law probability distribution: (1) increasing damage density with stress range S when $m > \alpha$, (2) decreasing damage density with stress range S when $m < \alpha$, and (3) a constant damage density with stress range S when $m = \alpha$. The implications of these observations are: for $m > \alpha$, the load data at the right rail dominates the accumulated damage, whereas for $m < \alpha$, the load data from the left tail are more damaging. These two cases represent the extreme/rare events at the tails. For $m = \alpha$, the damage contributions from all constituent loads are comparable, so that it belongs to the spectrum loads category.

3. CASE STUDIES AND THE RESULTS

3.1 Case-I: Vehicle road test score

The road test scores of more than 270 vehicles are collected and reported in the Consumer Reports' comprehensive test program [14]. The scores are presented on a 100-point scale. The tests were based on the results from more than 50 individual tests and evaluations, including performance, comfort and convenience, fuel economy, and more [14]. In performance assessment, the vehicles are divided by category and ranked according to their overall test scores. The test vehicles in terms of categories are: electric cars/plug-in hybrids, subcompact cars, small 2-door cars, compact cars, midsized cars, large cars, luxury compact cars, luxury convertibles, luxury midsized cars, ultra-luxury cars, sports cars, wagons (all-wheel drive), small SUVs, midsized SUVs, large SUV, luxury compact SUVs, pickup trucks etc. [14]. Some models are included in multiple categories, as appropriate. Fig.4 shows the histogram of the road test scores of all of the vehicles. A histogram divides sample values into many intervals called bins. Bars represent the number of observations falling with each bin (its frequency).

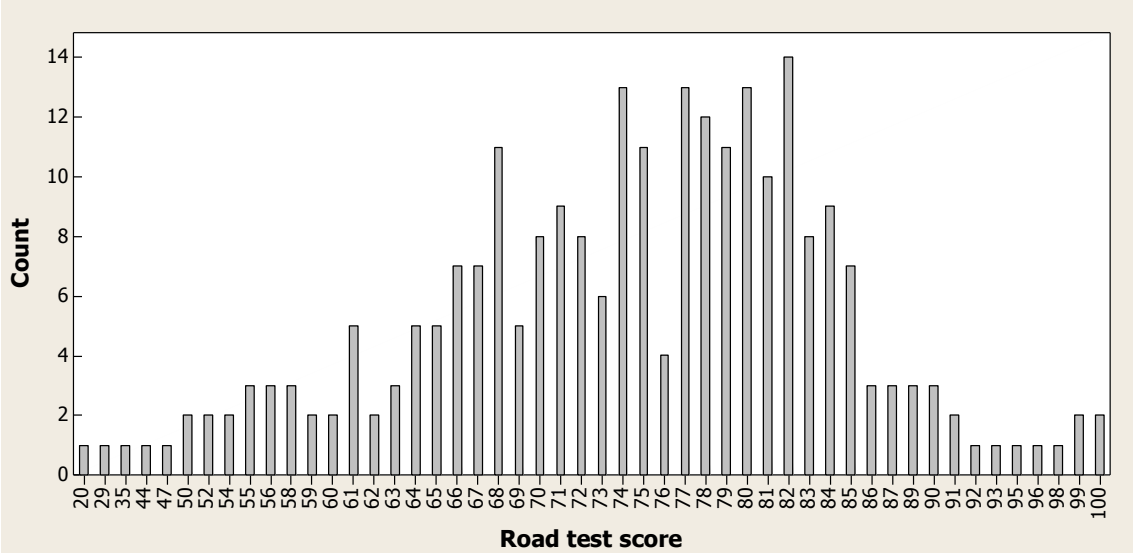
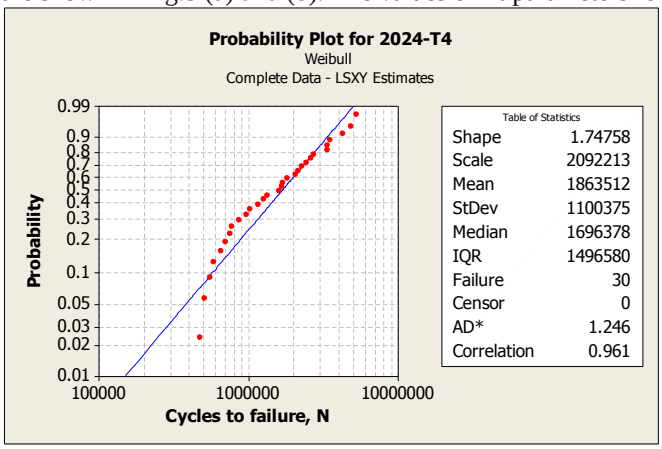


Fig.4 The counted number at each vehicle test score in a bar chart form

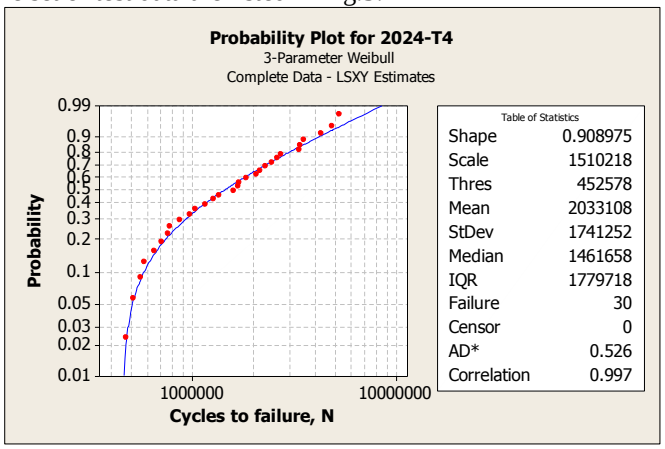
From Fig.4 it is seen that more vehicles are located in the score range [60, 85], resembling the normal distribution but the distribution is bounded at both right and left sides. It should be noted that each subcategory may show different patterns. However, this paper does not attempt to investigate the subcategories because of very limited vehicle number in each subcategory. Overall, Fig.4 provides the general features and patterns of the distribution of the vehicles as well as provides the bounds of the scores in the tested vehicles. The measurement system is bounded at 100 at the right side, while the rest data are tailing off to the left. The best vehicle according to the scores is Tesla Model S P85D (100) in both the categories of electric cars/plug-in hybrids as well as ultra-luxury cars, followed by BMW 750i xDrive (99), BMW M235i (98). The vehicle with the lowest score is Jeep Wrangler Limited (20), which is followed by Mitsubishi Mirage ES (29). Even though 20 is the lowest value in the ranking, it does not mean it is the lower bound in the distribution. First, the test score could be worse. Second, for a vehicle, it must satisfy some basic functions and therefore, it could not get a 0 in the score. This primarily depends on how the score system is pre-determined. Furthermore, no quantitative probabilistic fit is attempted for this set of data because the scores contain subjective measures, such as comfortableness, which is tester dependent.

3.2 Case-2: Fatigue data of 2024-T4

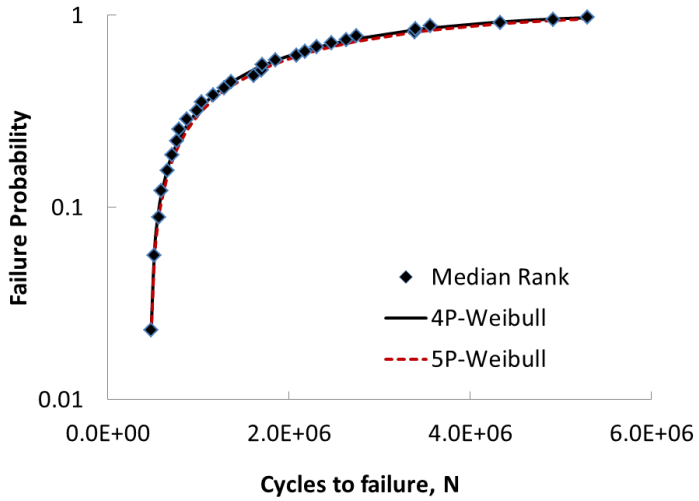
A set of high-cycle fatigue data at room temperature [1] with sample size of 30 for 2024-T4 is selected for fitting the Weibull distributions. The probability plots estimated using Minitab [15] for the two- and three- parameter Weibull functions are shown in Fig.5 (a) and (b). The values of fit parameters for the set of test data are listed in Fig.5.



(a)



(b)



(c)

Fig.5 Probability plots of (a) two-parameter Weibull distribution, (b) three-parameter Weibull distribution, and (c) four- and five-parameter Weibull distributions for a set of 2024-T4 data.

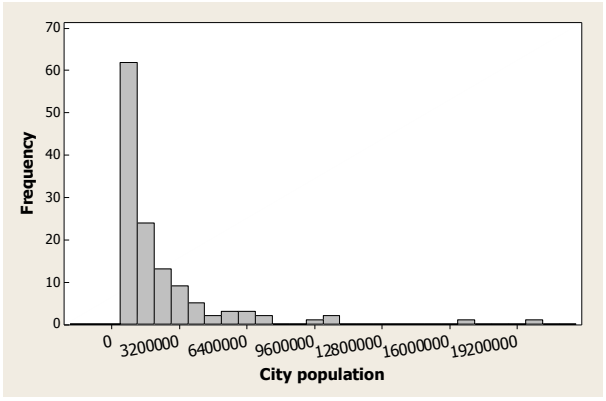
Table 1 The fit of 2P-, 3P-, 4P-, and 5P Weibull distribution functions for the high-cycle fatigue data of 2024-T4.

Distribution functions	Parameters	
2P-Weibull	β	1.74758
	η	2092213
	AD statistic	1.246
3P-Weibull	β	0.908975
	η	1510218
	δ	452578
	AD statistic	0.526
4P-Weibull	a	463000
	b	9760000
	β	0.77
	η	0.21
5P-Weibull	a	460000
	b	10000000
	β_1	0.80
	β_2	0.62
	λ	5.2

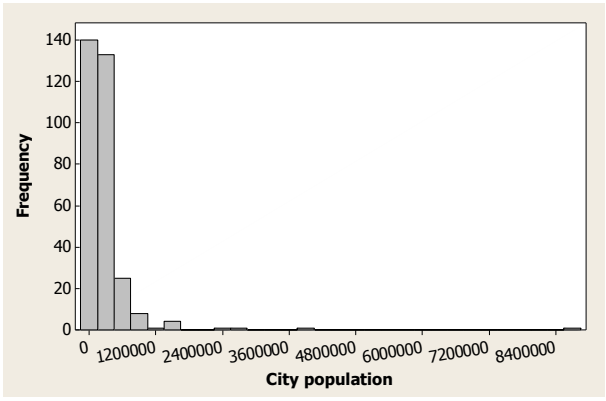
The data of cycles to failure show a large scatter. The three-parameter Weibull distribution has a much better fit in terms of visual examination and the AD statistic value. The AD values for the two- and the three-parameter Weibull distribution functions are, respectively, 1.246 and 0.526. The improved fit quality is particularly prominent at the left tail indicating the need of introducing the threshold parameter as the lower bound. From Fig.5 (b) it can be seen that even with 3-P Weibull distribution, the right side does not fit well and the predicted curve is below the data. For this data pattern, a more parameter controlling the local behavior of the data at the right side may be required. The probability plots calculated from the four- and five- parameter Weibull distributions are plotted in Fig.5(c). It is clear that the introduction of the fourth and the fifth parameters into the empirical Weibull distributions significantly increase the quality of fit. All the fitting parameters are listed in Table 1.

3.3 Case-III: City population

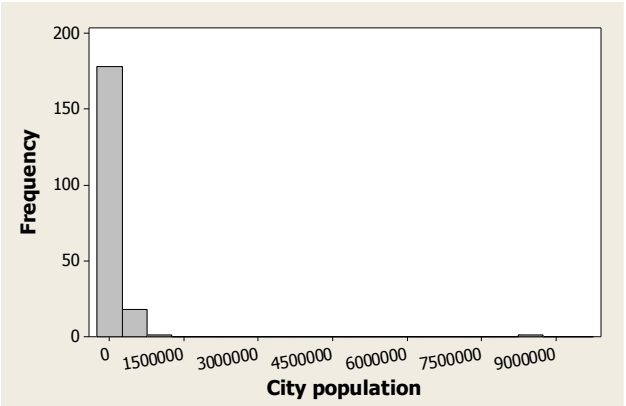
The population distribution of cities in 3 countries: China, the USA, and the UK, is analyzed here. The data are taken from the website www.citypopulation.de, which has compiled a large amount of data from various sources. The histograms of the city population of the 3 countries are shown in Fig. 6(a), (b), and (c), respectively. Clearly, the small cities are much more than larger cities. The city population in China was reported in November 1, 2010 and the largest city is Shanghai (20,217,748), followed by Beijing (16,446,857). The city population in the USA was reported in July 1, 2015, and the largest city is New York (8,009,185), followed by Los Angeles (3,485,398). The city population in the UK was reported in June 30, 2011, and the largest city is London (8,618,552), followed by Birmingham (1,115,791). In order to test if the Pareto power law distribution fits the data well, the complementary cumulative distribution function (C-CDF), Eq.(8), as a function of city population are shown in Fig.7 in a log-log plot. It is found that the data of the USA shows a decent linear behavior. By contrast, the city population of China and the UK in particular show a good linearity at small city population but deviate from the linearity at large city population. Overall, the Pareto law can be used to fit the data and the plots are shown in Fig.8. The median rank results obtained using Benard's approximation, Eq.(1), and the predicted results match very well. The values of the fit parameter α and the related information are listed on Table 2.



(a)



(b)



(c)

Fig.6 The histograms of the city population in (a) China, (b) the USA, and (c) the UK.

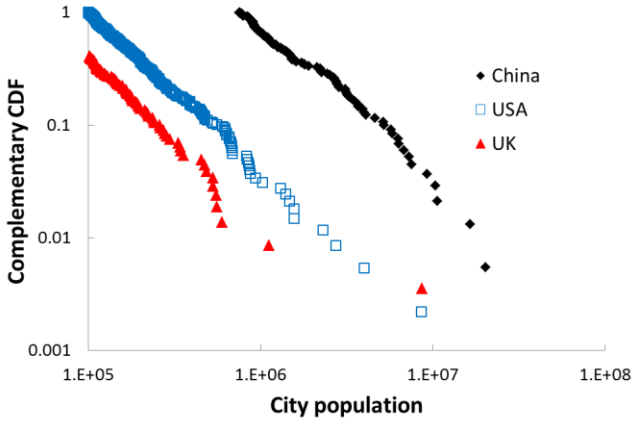


Fig.7 The complementary CDF as a function of city population in the three countries: China, the USA, and the UK.

Table 2 The population in the major cities in China, the USA, and the UK with power law fit.

Country	The number of cities, n	$x_{\min}$	α
China	128	750000	2.2595
USA	315	100000	2.4359
UK	198	50000	2.3690

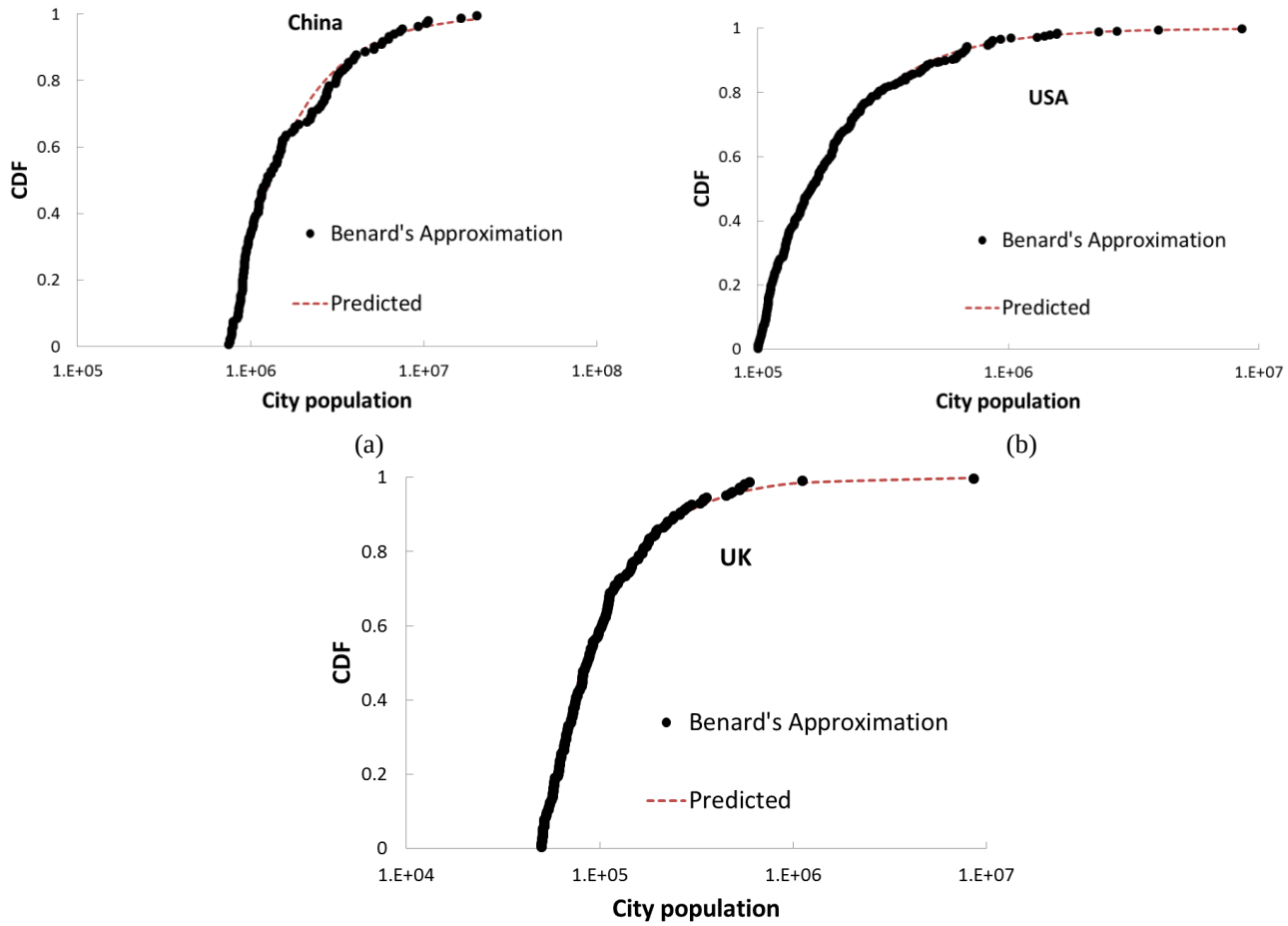
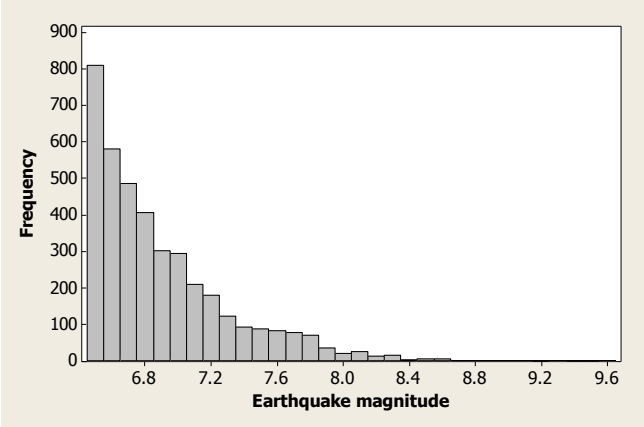


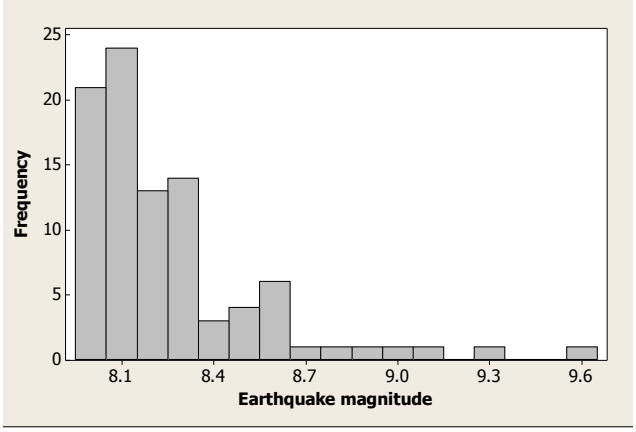
Fig.8 The CDF as a function of city population in the three countries: China, the USA, and the UK.

4.4 Case-IV: Earthquakes

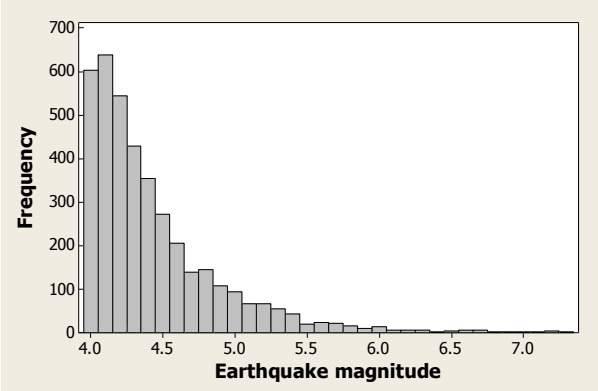
Worldwide earthquake with magnitude of 6.5 or above and the conterminous USA with magnitude of 4 or above are analyzed here. The conterminous U.S refers to a rectangular region including the lower 48 states and surrounding areas which are outside the U.S. The data are taken from the website http://earthquake.usgs.gov/earthquakes/. The worldwide earthquakes collected occurred between January 1, 1900 and July 17, 2016. The conterminous USA earthquake occurred between January 1, 1980 and July 17, 2016. The largest recorded earthquakes worldwide during this time period is the one that occurred in Chile in 1960 with magnitude of 9.6. The histograms of the two earthquakes are shown in Fig.9 (a), and (c), respectively. Fig.9 (b) is the same as Fig.9(c) but it starts from magnitude of 8.0, so that the events that are too small to be shown in Fig.9 (a) can be revealed. Similar to the city population, the small earthquakes occur much more than larger earthquakes. In order to test if the Pareto power law distribution fits the data well, the complementary cumulative distribution functions (C-CDF) as a function of earthquake magnitude are shown in Fig.10 in a log-log plot. It is found that the data of the worldwide earthquake shows a good linear behavior, but the conterminous USA shows a clear "bending down" phenomena. The Pareto law is used to fit the data and the plots are shown in Fig.11. The values of the fitting parameter α and related information are listed on Table 3.



(a)



(b)



(c)

Fig.9 The histograms of the earthquakes occurred (a) Worldwide with $M \geq 6.5$, (b) Worldwide with $M \geq 8.0$, and (c) the conterminous USA with $4 \geq 8.0$.

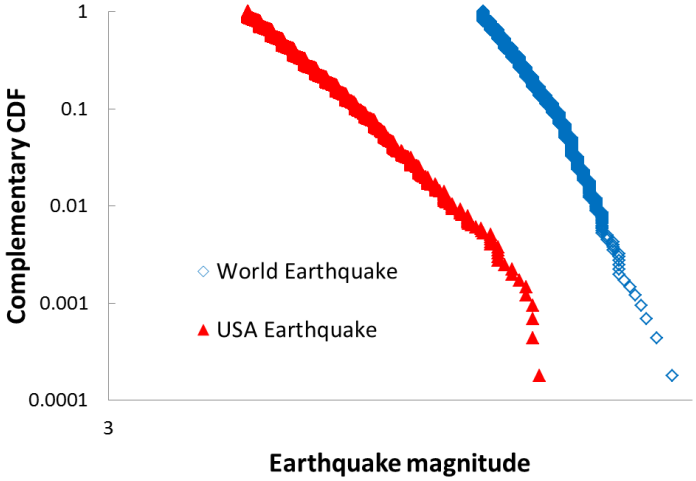
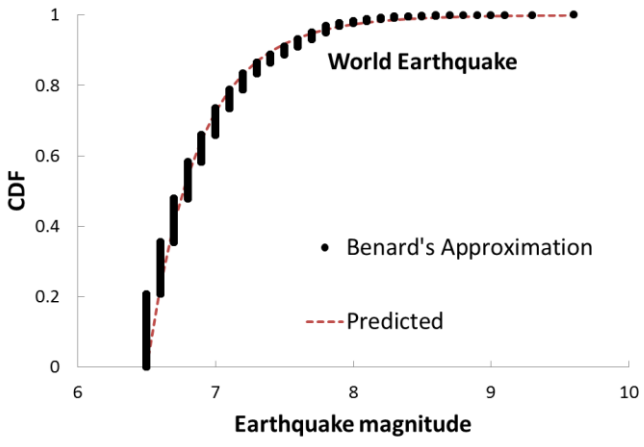
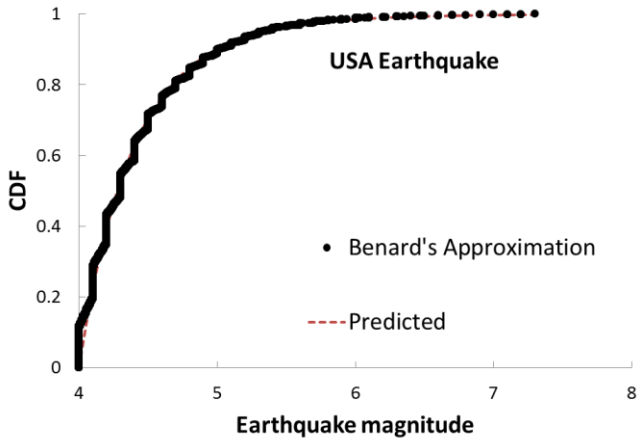


Fig.10 The complementary CDF as a function of earthquake.



(a)



(b)

Fig.11 The CDF as a function of earthquake magnitude of (a) Worldwide with $M \geq 6.5$, (b) the conterminous USA with $M \geq 4.0$.

Table 3 The fit values of parameter α for the earthquakes

Earthquake region	The number of earthquakes n	$x_{\min}$	α
Worldwide	3919	6.5	18.4903
Conterminous USA	3906	4.0	11.3662

4. DISCUSSION

The earthquake in conterminous USA and the city population in the UK and China show the "Dragon King" phenomena. Definitely, uncertainty is higher for extreme events such as mega cities and earthquakes. However, the uncertainty in the extreme large events cannot not completely conceal the general trends as shown in Fig,7 and Fig.10. In order to empirically describe the "Dragon King" phenomena the power law distribution should be improved by introducing more parameters, which could provide better fit to the observed data. The size of the mega city is much dependent on the location, climate, geopolitical condition, and policy. Global occurrence of magnitude 9 earthquakes is 1-3 per century [16, 17], so longer time is required to collect more reliable data, which limits the capability in prediction and forecasting of the extreme/rare events. Finally, the cumulative distribution function is not visually sensitive of the "Dragon King" events. This point can be clearly be reflected from the fact that the CDF generally provides satisfactory visual data fits for all of the data studied even though the C-CDF plots show clear "Dragon Kings" phenomena. Therefore, other measures such as the probabilistic density distribution should be used as a complementary tool.

5. CONCLUSIONS

1. A probabilistic event can be categorized as extreme/rare or spectrum events. Extreme/rare events are often located at the right tail of a probabilistic distribution.
2. The mean behavior of a spectrum event can be described using the conventional two-parameter distributions such as Weibul distribution; the four- and five-parameter Weibull functions provide the capabilities to accurately model the probabilistic distribution from the left tail to the right tail; the peak-over-threshold events can often be modeled with Pareto distribution.
3. Based on the combination of commonality/rarity and impact/effect, probabilistic events can be categorized into four events quadrants: (a) Quadrant-I: Rare-High Impact, (b) Quadrant-II: Rare-Low Impact, (c) Quadrant-III: Common-Low Impact, and (d) Quadrant-IV: Common-High Impact. The damage density parameter can be effectively used to differentiate the spectrum events and extreme/rare events.
4. Four case studies, i.e. vehicle road test score, fatigue life data, city population in three countries, and the earthquake worldwide and in conterminous USA are provided to demonstrate the concepts and procedures developed in this paper.

References

[1] Z. Wei, L. Gao, L. Luo, K. Nikbin, "Modeling and Analysis of Extreme/Rare Events and Spectrum Events", PVP2016-63918, *Proceedings of the ASME 2016 Pressure Vessels & Piping Division Conference*, July 17-21, 2016, Vancouver, BC, Canada (2016).

[2] NASA/SP-2011-3421, *Probabilistic Risk Assessment Procedures Guide for NASA Managers and Practitioners*, Second Edition, Editors: M. Stamatelatos and H. Dezfuli (2011).

[3] M.E.J. Newman, "Power Laws, Pareto Distributions and Zipf's Law", *Contemporary Physics*, **46**, 323 (2005).

[4] V. F. Pisarenko and D. Sornette, "Robust Statistical Tests of Dragon-Kings Beyond Power Law Distributions", *The European Physical Journal Special Topics*, **205**, 95 (2012).

[5] J. Janczura and R. Weron, "Black Swans or Dragon-Kings? A Simple Test for Deviation from The Power Law", *The European Physical Journal Special Topics*, **205**, 79 (2012).

[6] N. N. Taleb, *The Black Swan: The Impact of the Highly Improbable*, Penguin, London, UK (1997).

[7] A. Bernard and E. C. Bosi-Levenbach, "The Plotting of Observations on Probability Paper", *Stat. Neederlandica*, **7**, 163 (1953).

[8] A. Gudmundsson, "Elastic Energy Release in Great Earthquakes and Eruptions", *Frontiers in Earth Science*, **2**, Doi:10.3389/feart.2014.00010, 2014.

[9] J.A. Kies, "The Strength of Glass", Naval Research Lab, Report No.5093, Washington D.C. (1958).

[10] F. Smith and D. W. Hoeppner, "Use of The Four Parameter Weibull Function for Fitting Fatigue and Compliance Calibration Data", *Engineering Fracture Mechanics*, **36**, 173 (1990).

[11] K.K. Phani, "A New Modified Weibull Distribution Function", *J. Am. Ceram. Soc*., **70**,182 (1987).

[12]Wei, Z., Dong, P., Kurth, R., Gao, L., "Extreme Value Distribution Theory and Its Application in Durability Analyses", *Proceedings of the ASME 2012 Pressure Vessels & Piping Division Conference*, PVP2012-78232, July 15-19, 2012, Toronto, Canada (2012).
[13] S. Alfarano and T. Lux, "Extreme Value Theory as a Theoretical Background for Power Law Behavior", In Claudio Cioffi-Revilla, editor, *Power Laws in the Social Sciences: Discovering Complexity and Non-Equilibrium Dynamics in the Social Universe* (2004).
[14]Consumer Reports: New Cars, January (2016).
[15] B. Ryan, B. Joiner and J. Cryer, *Minitab Handbook : Updated for Release 16*, Sixth Edition, Boston, MA, USA, (2012).
[16] R. McCaffrey, "Global Frequency of Magnitude 9 Earthquakes", *Geology*, **36**, 263 (2008).
[17] R. McCaffrey, "The Next Great Earthquake", *Science*, **315**, 1675 (2007).
[18] R.B. Bergmann and A. Bill, "On the Origin of Logarithmic-normal Distributions: An Analytical Derivation, and Its Application to Nucleation and Growth Processes", *Journal of Crystal Growth*, **310**, 3135 (2008).
[19] W. K. Brown and K. H. Wohletz, "Derivation of The Weibull Distribution Based on Physical Principles and Its Connection to The Rossin-Rammler and Lognormal Distributions", *Journal of Applied Physics*, **78**, 2758 (1995).
[20]W. Weibull, "A Statistical Distribution Function of Wide Applicability", *Journal of Applied Mechanics*, **18**, 293 (1951).
[21]M. Fréchet, "Sur la loi de probabilité de l'écart maximum", *Annales de la Société Polonaise de Mathematique*, **6**, 93 (1927).
[22] Z. Wei, F. Yang, S. Lin, D. G. Harlow, "Failure Modes Analysis of Fatigue S-N Test Data with Small Sample Size", RQD18-310, 18th ISSAT International Conference on Reliability and Quality in Design (18th ISSAT RQD), July 26-28, 2012, Boston, MA, USA.
[23] I. Elishakoff, *Probabilistic Theory of Structures*, Second Edition, Dover Publications, Inc., Mineola, New York (1999).
[24]Z. Wei, D.G. Harlow, S. Lin, F. Yang, "The Underlying Driving Forces of Continuous Probability Density and Distribution Functions", *18th ISSAT International Conference on Reliability and Quality in Design (18th ISSAT RQD), RQD18-309*, July 26-28, 2012, Boston, MA, USA (2012).
[25]Z. Wei, D. G. Harlow, K. Nikbin, "The Underlying Driving Forces of Multiple-modal Statistical Behavior", *Proceedings of the ASME 2014 Pressure Vessels & Piping Division Conference, PVP2014-28880*, July 20-24, 2014, Anaheim, California, USA (2014).

Appendices: Two underlying mechanisms for the Pareto power law distribution

Appendix-I: Equilibrium based distribution

Many attempts have been made to find the underlying physical bases of some specific probabilistic distribution functions. For example, the "bean machine" or "Galton board" is believed to be the first physical model that can generate the normal distribution [18]. Many efforts have been made to reveal the underlying mechanisms of normal, lognormal, and Weibull distribution functions [18-21]. The ubiquity of power-law behavior in the natural world has led many scientists to wonder whether there is a single, simple, underlying mechanism linking all these different systems together. Several candidates for such mechanisms have been proposed, going by names like "self-organized criticality" and "highly optimized tolerance" [3]. However, it should be emphasized that all of the models described above are based on a specific distribution function with a specific physical process such as crack initiation and growth or fragmentation.

Recently, a new approach [22] has been developed and all of the available probability density and distribution functions can be conceptually considered as a result of the equilibrium process from two competing driving forces: a short-range repulsive force and a long-range attractive force with proper boundary conditions and other constraints imposed. With this concept, a particular driving force model has been developed below: an elastic spring (inverse spring) model is selected as the short-range force, which can be applied only to the neighboring masses; a body force, exerted on a mass point with x as the coordinate, is selected as the long-range force with the attractor located in x_0 of the field. Fig.A1-1 schematically illustrates the model and the equilibrium mechanism. In real applications $p(x)$ represents a probability density function and x represents a variable of interest such as cycle to failure in fatigue tests. The equilibrium state can therefore be uniquely obtained by solving the following system of several governing equations: (1) a conservation equation, Eq. (A1-1), (2) a constitutive equation, Eq.(A1-2), and (3) an equilibrium equation, Eq.(A1-3). With these three equations the long-range force, Eq.(A1-6), can be determined if other conditions are provided or a probability density or distribution function, Eq.(A1-7), can be established when the two forces are given.

Conservation equation

$$\int_a^b n(x)dx = N \qquad (A1\text{-}1)_1$$

or

$$\int_a^b p(x)dx = 1 \qquad \text{(A1-1)}_2$$

Where $-\infty \le a \le x \le b < \infty$ and $p(x) = n(x)/N$, $n(x)$ can be considered as the number of mass points per unit length and N is the total points.

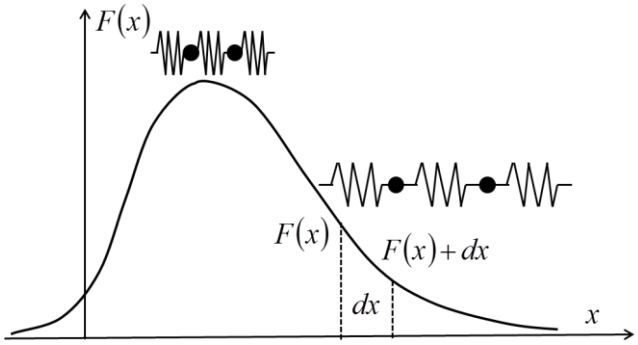


Fig.A1-1 Equilibrium analogy of distribution functions

Constitutive equation

$$F(x) = 1/L(x) \qquad \text{(A1-2)}$$

where $L(x)$ is the distance between the two adjoining masses and K is the stiffness of the spring and can be considered as a constant. The important characteristics of the spring model is $F \to 0$ as $L(x) \to \infty$ and $F \to \infty$ as $L(x) \to 0$. Therefore, this elastic model can be called the reverse spring model, which is different from the conventional spring model, i.e. $F(x) = KL(x)$.

Equilibrium equation

With the force analysis of a segment length dx, Fig.A1-1, we have the following force equilibrium equation

$$\left[F(x) + \frac{dF(x)}{dx}dx\right] - F(x) + f(x)n(x)dx = 0 \qquad \text{(A1-3)}_1$$

or

$$\frac{dF(x)}{dx} + f(x)n(x) = 0 \qquad \text{(A1-3)}_2$$

where $f(x)$ is the body force caused by the long-range force exerted on each small mass.

With the constitutive equation, Eq.(3), and the fact that

$$\frac{1}{L(x)} \propto n(x) \quad \text{or} \quad \frac{1}{L(x)} \propto Mn(x) \qquad \text{(A1-4)}$$

and then substituting Eq. (A1-4) into Eq. (A1-3)$_2$, we have

$$D\frac{dp(x)}{dx} + f(x)p(x) = 0 \qquad \text{(A1-5)}$$

Where M and $D = KM$ are constants.

Therefore, from Eq. (A1-5) the long-range body force can be expressed as Eq.(A1-6).

$$f(x) = -D\frac{dp(x)}{dx}\Big/p(x) = 0 \qquad \text{(A1-6)}$$

the minus sign here represents the force direction.

If both short-range and long-range forces are given, the probability density functions can be derived from Eq. (A1-6) as

$$p(x) = \exp\left[-\frac{1}{D}\int_a^x f(x')dx' + C\right] \qquad \text{(A1-7)}$$

where C is a constant, which should be determined by applied constraints such as given boundary conditions and conservation requirement. As a specific example, the long-range forces derived from Eq. (A1-6) for Pareto power law distribution, Eq.(6), is $f(x) = D\alpha x^{-1}$.

Appendix-II: Evolution based probabilistic distribution

A specific distribution function of an event such as the city population and cycles to failure can be treated as a result of evolution from a previous state by following a given evolution law. Mathematically, the problem is equivalent to seeking a target distribution function $P_Y(y)$ for a given initial distribution function $P_X(x)$ and a transformation function $y = \phi(x)$. The procedure [23] is well developed and has been applied to several applications [24, 25] and is briefly described below.

The target distribution function $P_Y(y)$ can be expressed as

$$P_Y(y) = P[\varphi(X) \le y] \tag{A2-1}$$

No matter, $y = \phi(x)$ is a strictly monotone increasing or decreasing function, $x = \psi(y)$ is a unique inverse function and

$$p_Y(y) = \frac{dP_Y(y)}{dy}\, p_X[\psi(y)]\left|\frac{d\psi(y)}{dy}\right| \tag{A2-2}$$

Fig. A2-1 schematically illustrates the process of function evolution. With the variable transformation technique shown in Eq.(A2-2), the distribution function $p[x(T=1)]$ at dimensionless time $T=1$ is the result evolved from a distribution function $p[x(T<1)]$ at a previous dimensionless time $T<1$ under a given evolution law. The process is reversible.

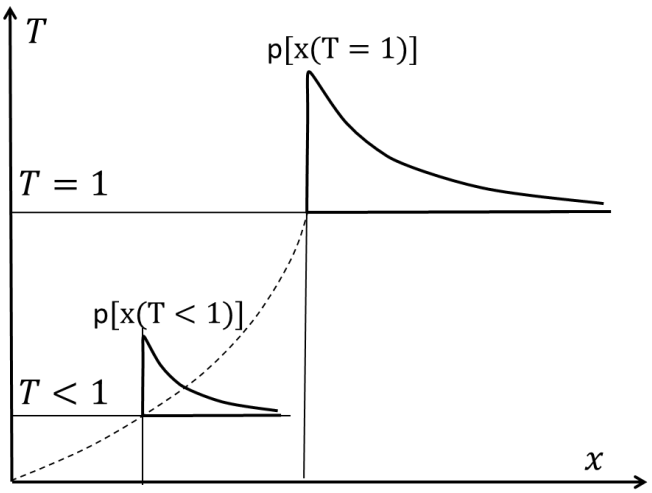


Fig.A2-1 Evolution of an exponential/power law type distribution function

An example is given here to illustrate the process. Assume a linear evolution law exists, and the relationship between the value of the independent variable at a historical point (x_H) and that at current point x is $x_H = \varphi(T) = Tx$ or $x = \psi(N) = x_H/T$. Then we have $d\psi(x_H)/dx_H = 1/T$. Finally, according to Eq. (A2-2) the distribution function at a historical point x_H can be derived as Eq.(A2-3).

$$f(x_H) = \frac{\alpha - 1}{Tx_{\min}}\left(\frac{x_H}{Tx_{\min}}\right)^{-\alpha} \tag{A2-3)$_1$}$$

$$F(x_H) = 1 - \left(\frac{Tx_{\min}}{x_H}\right)^{\alpha-1} \tag{`(A2-3)$_2$}$$

When the dimensionless parameter $T=1$, Eq.(A2-3) is recovered to Eq.(6) in the main text. Eq.(A2-3) indicates that for the linear evolution law, the distribution function is still a Pareto power law function with a lower bound parameter of $Tx_{\min}$.